DOCUMENT RESUME

ED 393 883 TM 024 785

AUTHOR Rasor, Richard A.; Barr, James
TITLE Refinement in Assessment Validation: Technicalities of Dealing with Low Correlations and Instructor Grading Variation.
PUB DATE [95]
NOTE 23p.; Portions of paper presented at the Annual Research Conference of the RP Group Granlibakken, 1993.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Students; *Correlation; Educational Assessment; *Grade Point Average; *Grading; Higher Education; *Scores; Student Evaluation; Test Results; *Validity

ABSTRACT

Issues and problems in assessment research are explored, with suggestions to help establish an acceptable correlation between student assessment scores and final grades. Topics include assessing instructor grading variation, instructor grade point average (GPA), success rates, and lack of linearity in grade scales. Solutions to these problems are offered in the form of a new four-point research grading scale and a new "contextual" student GPA based on cumulative grade average with the final grade in the target course removed from the calculation. The contextual GPA was used to equate student "skill" levels in course selections when identifying the degree of instructor grading variation. The techniques were applied to a sample of 6,077 students covering performance data in 26 courses. The magnitude of the resulting correlations suggests giving much greater emphasis to student cumulative college GPA as a multiple measure when establishing entrance "skill" levels deemed necessary for success in general courses having no specific course prerequisite. Practical suggestions are included for identifying the true correlation between assessment test scores and grades given the problem of instructor grading variation. (Contains three tables and six figures.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 393 883

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RICHARD A. RASOR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

REFINEMENT IN ASSESSMENT VALIDATION:

TECHNICALITIES OF DEALING WITH LOW CORRELATIONS AND INSTRUCTOR GRADING VARIATION

by

Richard A. Rasor

&

James Barr

American River College

TM024785

BEST COPY AVAILABLE

Abstract

Portions of this paper were originally presented at The RP Group's 1993 Annual Research Conference at Granlibakken. The report covers many issues and problems in assessment research along with suggestions that may prove helpful in establishing an acceptable correlation between student assessment scores and final grades.

Topics include assessing instructor grading variation, instructor GPAs, success rates, and lack of linearity in grade scales. Solutions to these problems are offered in the form of a new 4-pt. research grading scale, and a new "contextual" student GPA based upon cumulative grade average with the final grade in the target course removed from the calculation. The contextual GPA was used to equate student "skill" levels in course sections when identifying the degree of instructor grading variation.

The authors applied their techniques to a large sample of students covering performance data in several general education courses. The magnitude of the resulting correlations suggests giving much greater emphasis to student cumulative college GPA as a multiple measure when establishing entrance "skill" levels deemed necessary for success in general courses having no specific course prerequisite.

The report ends with practical suggestions on how to identify the true correlation between assessment test scores and grades given the problem of instructor grading variation.

Acknowledgments

We wish to thank Sharon McCuen, Dean of Research and Development at American River College, for her continued efforts in promoting institutional research.

Dr. Queen Randall, former president of American River College and now Chancellor, Los Rios Community College District, is also thanked for her invaluable support regarding campus based research.

Finally, we would be unable to conduct our present line of research without the continued efforts of the fine research staff at the Los Rios Community College District Office, namely, Jose Pagtalunan, Brad Brazil, and Janice Jones (presently Interim Dean of Math and Engineering, American River College).

Many community college researchers, assessment personnel, and various staff concerned with matriculation throughout California, are still having a difficult time with validating assessment/course placement procedures at their colleges. It would seem to many of these people that establishing a correlation coefficient of at least .35 (which is state mandated) between student assessment scores and course grades is like asking for the moon! However, the problem is not that .35 is too high a standard. Few of us would expect a respected test publisher to ever release an instrument with validity coefficients any lower. Indeed, we would probably demand much more. Problems arise when we apply validity standards in a micro sense, that is, to one type of course, often with many levels, with different instructors, and usually with small sample sizes.

The difficulties in not finding a correlation of .35 or higher between assessment test scores and course performance stem from several reasons:

Possible Causes of Low Correlations

1. A test publisher may validate a test in a general way using somewhat different procedures and applying different norms than what we do in our attempts at local micro validation.

2. Usually test publishers do not have to validate scores against several levels of one type of course, nor do they have to deal with intact systems where there is apt to be considerable resistance toward temporarily suspending rules of placement for the sake of validating an existing (and perhaps long-standing) placement test.

3. Single test validation may not be uniform for all colleges within a multi-college district because each institution has its unique characteristics including different student populations.

4. Within certain courses, there appears to be considerable grading variation between instructors who teach the same course. This means that the traditional criterion of final grade can be highly unreliable. If one instructor's "A" represents the same degree of excellence as another instructor's "C", then there can only be a low correlation between assessment test scores and grades.

5. Final grades as a criterion measure are often based upon course completers. Yet such students may represent a restricted range of talent when compared to the original class. This is because the students who dropped out could have had lower assessment test scores. Any restriction of range (predictor or criterion measure) are apt to lower a correlation.

6. There appears to be no clear difference between the letter grade of "F" and the designation "W" (indicating withdrawal from the course). In a recent survey at ARC, 67% of the teaching faculty routinely drop a student who is no longer

attending class while nearly 32% give the "F" grade (1% give an Incomplete). This means that how one codes such grades for computational purposes has a substantial impact upon the value of any correlation between assessment test scores and grades.

7. From ARC's experience, "A", "B", or "C" final grades are linearly related to assessment test scores (linearity is a prerequisite for using the Pearson correlation coefficient). However, grades of "D", "F", or "W" are sometimes associated with high assessment test scores. In one of our earlier research projects, students who withdrew from English 1A, as a group, had the highest assessment test score average. Such lack of linearity between measures lowers any correlation.

8. Dated assessment test scores (over six months old) with subsequent enrollment in other courses are apt to lower a correlation between scores and target course grades. Yet the reality is that enrollments in courses affected by matriculation regulations include many experienced students who enroll in the target course long after an assessment test was taken.

9. Course placement recommendations given to students in private by counseling staff (or through the student "grapevine") may sometimes include a subtle suggestion that a particular student who scores low on an assessment test take a particular instructor that has a reputation for "easy" grading. The converse may also be true. Such practices have resulted in low assessment scores being paired with high grades, an outcome sure to lower any overall positive correlation.

10. In spite of the previously mentioned reasons (and probably many others), there remains the distinct possibility that little or no correlation exists between assessment scores and grades in a target course. Without a task analysis done on the course, content analysis done on the assessment test, and determining how grades are assigned, one cannot be sure even why there <u>should</u> be any relationship.

We address some of these technical problems which you may find useful in your assessment research.

<u>Instructor Grading Variation, the Conventional Course GPA, and the Success Rate</u>

In examining the GPAs for entire sections of courses (number of "A"s, "B"s, "C"s and other grades given, it became obvious to us that instructors were not applying grading standards in the same way. The worst example we found at ARC was a course GPA of 1.28 for 10 sections with one instructor (and a success rate of 32%), while for a second instructor of the same course, the GPA was 3.01 and based upon 15 sections (with a success rate of 81%). Remember that the dropout rate is not included in course GPA (nor a student's conventional GPA) but is included in the computation of success rate. Clearly, in this instance, which instructor a student enrolled with probably had much more to do with course success than did any assessment test score.

In terms of establishing validity coefficients, we do not like the conventional GPA. The reason is simple enough, GPA does not include "W" notations. An instructor could give one "A" and have the rest of the class drop which would result in a course GPA of 4.0!

At first consideration, course success rates would seem to be a better index of overall class performance because it is easily understood and includes the number of students receiving "W" notations. But, it too, has its limitations. With identical success rates of 50%, one instructor could have assigned half the class "C"s (with the other half receiving "W"s) while another instructor could give all "A"s to half the class (with the remaining half receiving "F"s). What is needed is a modified success index that is sensitive to degrees of success as is the conventional GPA. A practical solution will be discussed shortly.

Assigning Values for the Coding of Letter Grades

Our Assessment Director (Tom Powell) had already run many correlations between assessment scores and grades in selected courses. His printouts usually included two correlation values, one with "W"s coded the same as "F"s (both zero), and one with all "W"s removed from the calculations (which lowered sample size).

We examined the mean assessment score for students who had received a specific grade in a target course (i.e., assessment mean cross- tabulated with letter grade level). We did this for each grade level in several courses. Our results are presented in Table 1 (p. 4), and are based upon 26 courses and 6,077 students. In 24 of 26 courses, the "A" students had the highest assessment score mean (an assigned rank order of 1). In 2 of the 26 courses, the students who earned an "A" grade had an assessment score mean that fell in 2nd place (below some other group of students who earned a different grade). Clearly then, students who earned an "A" in nearly all of these courses also had the highest assessment scores when compared with students who earned different letter grades. So far, this speaks well for linear trend. However, subsequent grades do not follow that linear pattern so nicely. In other words, students who earn "B"s have assessment score means which are not always in second place standing (rank order = 2). And so it goes.

When the median rank orders are computed and plotted along a straight line, an interesting finding emerges . Students who earn an "A" grade, as a group, usually have the highest assessment score mean irrespective of course. The "B" group usually comes in 2nd place with respect to their assessment mean. With "C"s, the rank orders are much more varied (i.e., three times in 2nd place, eight times in 3rd place, -- even two times in 6th place below "A", "B", "D", "F", or "W"). Clearly, linearity breaks down starting with the letter grade "C", then dramatically so after the "C". Grade groupings of "D", "F", and "W" are so mixed that their median ranks based upon assessment test means are nearly identical.

Table 1

Assessment Test Score Means Rank-Ordered For Students Earning

A Specific Final Grade [1]

Grade	Rank 1st	Rank 2nd	Rank 3rd	Rank 4th	Rank 5th	Rank 6th
A	24	2				
B	1	19	6			
C		3	8	9	4	2
D	1		6	5	7	7
F		1	4	6	7	8
W		1	2	6	9	8

[1]These results represent the number of instances (out of 26 courses) that the assessment mean score for students receiving a particular grade fell at an ordinal position of six possible grade rankings. The data are based upon 26 courses and 6,077 students who took either the APS test for placement in general education courses, or the MDTP test for math placement. For example, all students receiving an "A" grade in a particular course had assessment test score mean that was the highest (rank = 1) in 24 out of 26 instances. In two instances the "A" grade group had an assessment test mean that fell second highest (rank = 2). The median ranks for each grade level across all courses are:

A = 1.04, B = 2.13, C = 3.72, D = 4.64, F = 4.79, W = 4.94

Given the finding that the "D" "F" "W" grades represent nearly interchangeable rank ordered values, we recoded letter grades and recomputed the correlations between the test scores and grades for all 26 courses. The recoding was done as follows: "A" = 4, "B" = 3, "C" + "CR" = 2, and "D" or "F" or "NC" or "W" = 1 (incompletes or in-progress grades were left out because they are temporary notations). By this simple recode, we found that 76% of our correlations increased in magnitude over the original values. This indicated to us that part of the low correlation problem is lack of a linear relationship between assessment test scores and grades. The problem can be rectified somewhat by the recoding as suggested.

BEST COPY AVAILABLE

Development of the 4-Point Research GPA

On page 3 we indicated that a modified success rate that would be sensitive to "degrees" of success was needed. Our recoding of grades as "A" = 4, "B" = 3, "C or CR" = 2, and all other unsuccessful grade notations = 1, had worked out well for our correlational assessment research. We also thought that the same recode should be useful with research on instructor grading variation. In other words, instead of comparing conventional instructor GPAs or simple success rates, why not calculate this new GPA? (We dubbed this a "research GPA" to distinguish it from the regular GPA). The benefits include relative ease of computer recoding, the resulting mean (GPA) closely approximates the mean for the regular GPA, the scale overcomes trying to distinguish between "F"s and "W"s, and because of including "W"s, helps to maintain a desirable sample size. With the 4-point research GPA, a value of 2.0 would mean that all students averaged out at a "C" level. A value of 1.0 would be interpreted as all students averaging out as unsuccessful ("D" or "F" or "NC" or "W").

As a pilot test, we constructed 50 hypothetical instructor grade distributions and calculated the correlations between research GPA and simple success rate (r = .93), between research GPA and regular GPA (r = .93), and between regular GPA and success rate (r = .87). These preliminary results were encouraging.

The merit of the research GPA is that it maintains the virtues of the success rate without sacrificing the power of detecting subtle grade differences within the definition of success ("A", "B", "C", "CR"). The drawback is that reporting another type of GPA may prove confusing. One possible solution to any confusion would be to multiply the research GPA by 100 and round the value. You could call this a course performance score.

Consider the following hypothetical grade distribution for one instructor in a specific course:

Grade	frequency	Gradepoints
"A"(4)	20	80
"B"(3)	30	90
"C"(2)	10	20
"D"(1)	5	5
"F"(1)	15	15
"W"(1)	20	20
Sums	100	230

Regular GPA = 2.44
Success Rate = 60%
Research GPA = 2.30
Course Performance Score = 2.30 x 100 = 230 (out of 400 possible)

The course performance score can be based upon any grade coding. In this example, the score of 230 indicates that the overall class reached a little better than "C" (or 200) on the 4-point scale.

Intercorrelations With The Research GPA

Previously, we indicated that the research GPA was highly correlated with success rate. But the correlation of .93 was based upon 50 hypothetical grade distributions. Not being entirely content with that approach, we also constructed a research GPA, a regular GPA, and a success rate from the cumulative college records for each of 6,955 students who were presently enrolled in at least one of 19 different general education courses. The intercorrelational scatterplots are presented as Figures 1, 2, and 3 (see pages 7,8,9).

In Figure 1, the correlation between the 4-point research GPA and the success rate (expressed as a percent) is .90. Notice at the top of the plot how many students have 100% success rates but research GPAs falling between 2.0 and 4.0. It should be pointed out that there were 550 students in our sample who had a research GPA of 0.0 and a success rate of 0%. When these students were temporarily deleted from the computation, the correlation was reduced to .87 which is still a strong relationship.

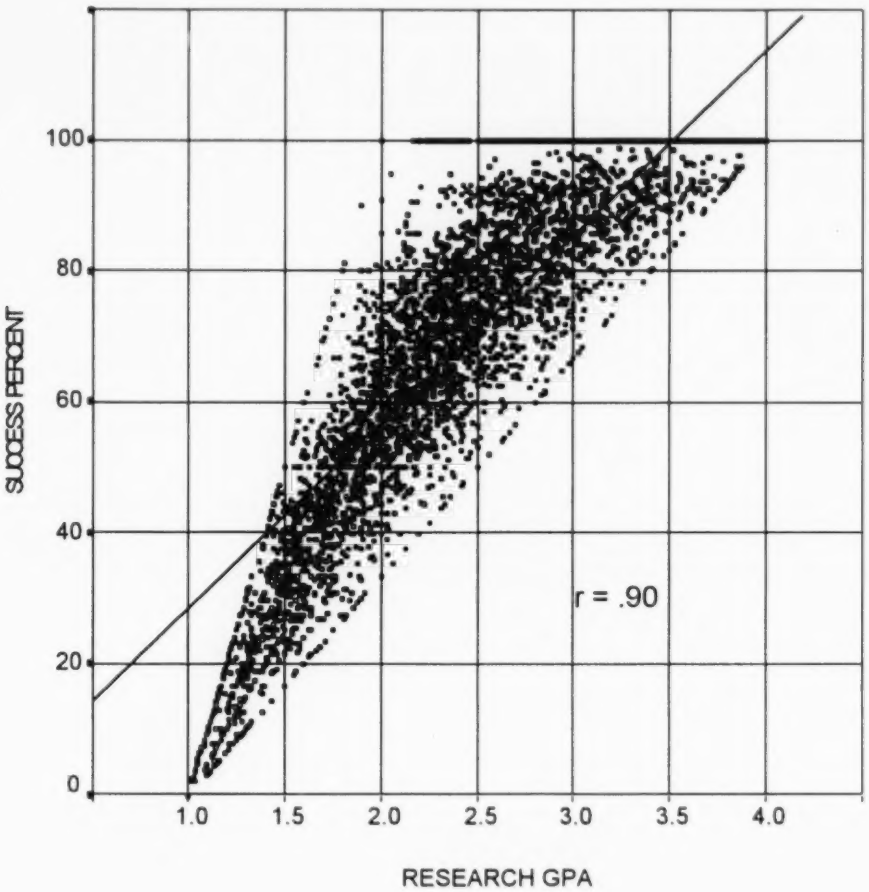

Figure 1. Scatterplot between research GPA and percent of success for 6,955 students.

BEST COPY AVAILABLE

We also wanted the research GPA to be highly correlated with regular GPA. Figure 2 depicts the relationship which is also positive and strong (r = .90). (Note: Deleting the 550 students mentioned previously resulted in a correlation of .89). In Figure 2, notice the data points representing regular GPAs of 4.0 (straight "A"s) but with corresponding research GPA's falling between 1.0 and 3.0. These are students who either dropped most of their classes or got "NC" notations which are not included in regular GPA.

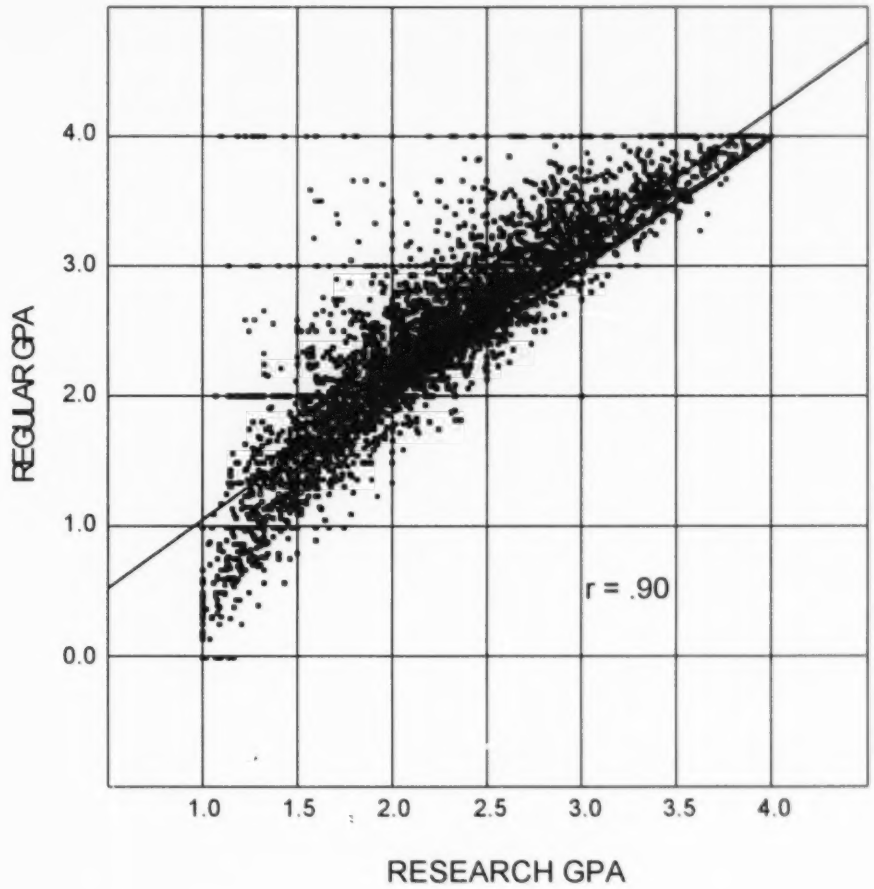

Figure 2. Scatterplot between regular GPA and research GPA for 6,955 students.

Figure 3 depicts the relationship between regular GPA and success rate (r = .84). Once again, the relationship is both positive and strong but a little lower than the other correlations. Also notice the students who had perfect 4.0 GPAs but low success rates (representing dropping all courses or receiving "NCs" except for those in which an "A" was earned). Note: When the 550 students were deleted from this computation, the Pearson r was reduced to .75.

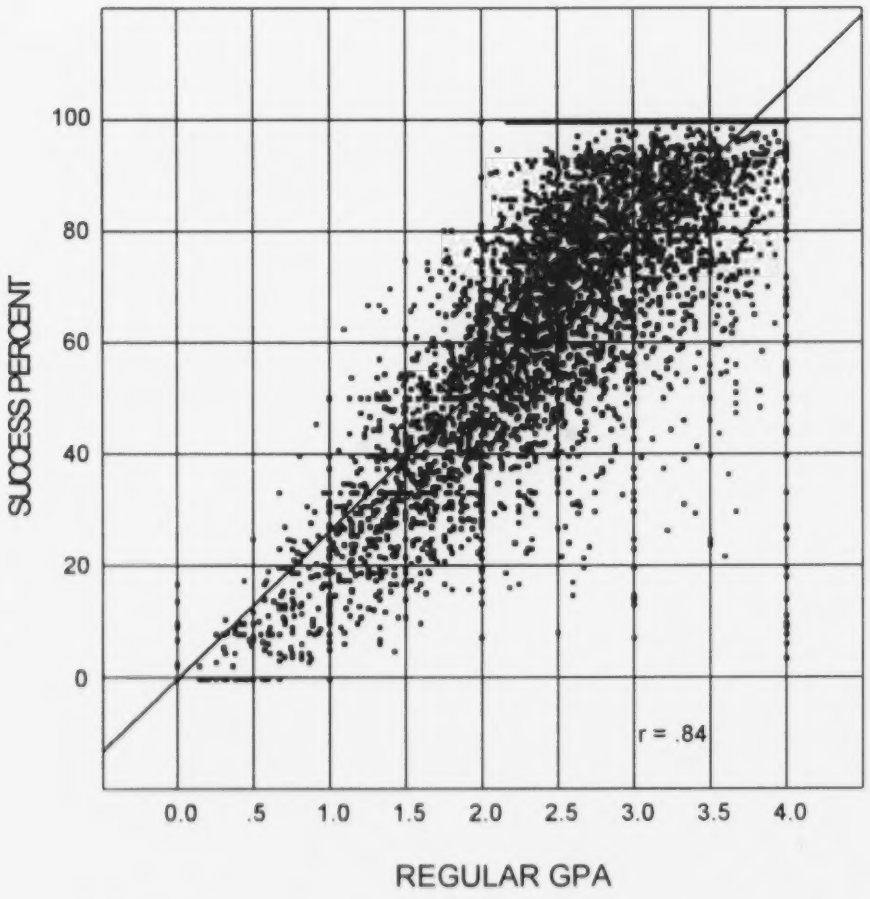

Figure 3. Scatterplot between percent of success and regular GPA for 6,955 students.

Shape of the Distribution for the Research GPA

The frequency distribution for the research GPA is presented as a histogram in Figure 4. The elevated bar on top of 1.00 indicates that there were 550 students who were totally unsuccessful in terms of grades (about 8%). The mean is 2.30 which compares nicely with 2.40 for a regular GPA. Apart from the elevated bar above 1.0, there is a subtle positive skewness (tail more on the right side).

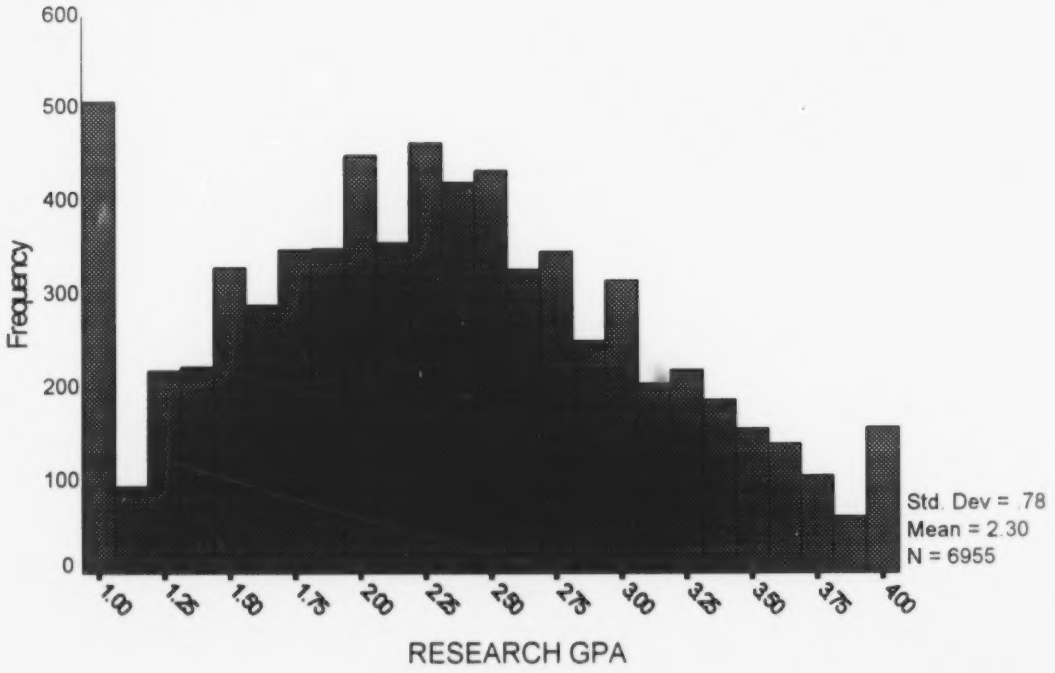

Figure 4. Histogram showing frequency of research GPA.

Figure 5 represents the same students and their regular GPAs (which omits all "W" or "NC" notations). This distribution is negatively skewed (tail on the left side of the curve) and presents a much more favorable picture of student performance. At this point the reader may wonder why anyone would want to use the research GPA when the regular GPA presents findings in a more favorable light? Our answer is that we believe the 4-point research GPA more accurately reflects the students' total academic performance. It is for this reason, plus the fact that research GPA is highly correlated with success rate, that we plan to use it in most of our research inquiries that involve student performance measures. However, we are not likely to change the official system of grading, so we recommend restricting the use of research GPA to just that, research.

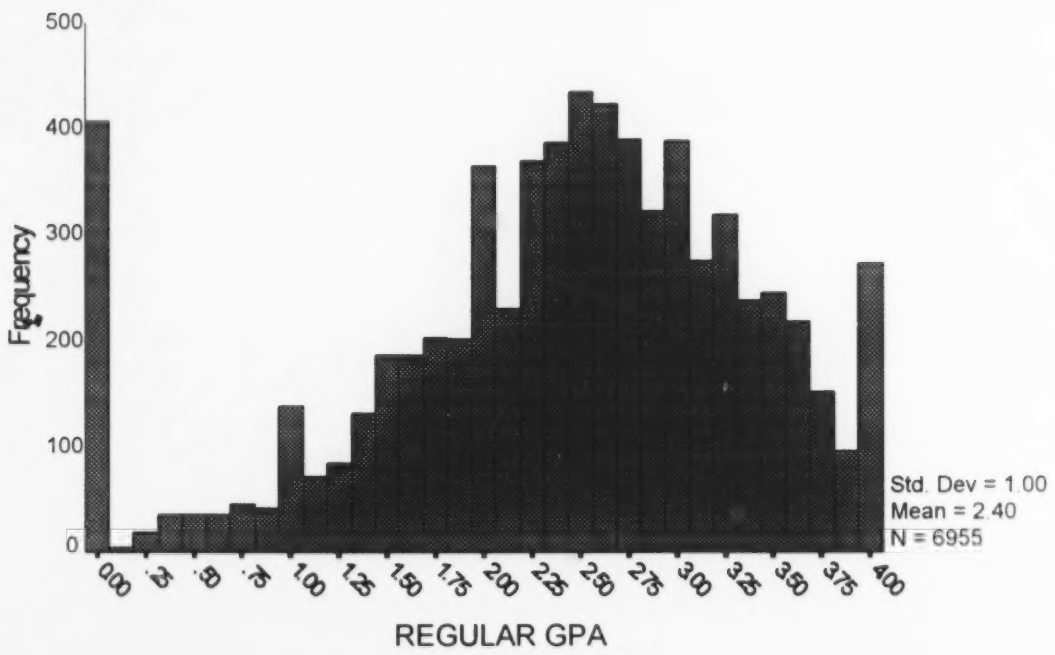

Figure 5. Histogram showing frequency of regular GPA.

Development of the Contextual GPA[1]

In planning research studies on the subject of instructor grading variation, one always has to wonder if students enrolled in the same course but with different instructors and/or different sections, average out at the same ability level? In other

[1] The computer programming for any of the GPAs in this report can be obtained by writing to Jim Barr, c/o ARC.

words, could a substantial amount of differences in instructor grading (or retention) be due to initial differences in student ability/motivational levels? We used to think that our registration process was akin to random assignment of students to course sections thereby balancing out student skill levels across instructors. But then our subsequent research with English 1A instructors revealed a relationship between skill level and instructor selected. We found that lower assessment scoring students enrolled in English 1A with instructors who gave the largest number of high grades! (This really should not have been a surprise but it was). Furthermore, we also discovered that students who had the highest assessment score mean were enrolled with an instructor who typically gives low grades!

This inverse relationship indicated to us that in order to do good research on instructor grading variation we must control for the variable of student ability. Unfortunately, the relationship between scores on our assessment instrument (APS) and grades in English is quite low, and some students do not have a test score on file or the score is considerably dated. Furthermore, new freshmen have no college GPA nor are many high school transcripts available when they could be useful. What we needed was a performance measure on every student that was timely and that would be independent of the grade earned in the target course.

Our solution was to develop a 4-point research GPA, as before, on each student through enrollment in the target course, but with this important modification: The grade in the target course would be removed from calculation of the research GPA. We dubbed this value as the "contextual GPA", sort of a wrap-around-the-target-course research GPA.

To illustrate, assume that a student took nothing but 3-unit courses and the following grade notations (minus the target course grade) appeared on her cumulative record:

AAAABBBCCCFWW plus earned a "C" in the target course.

Grade	frequency	Gradepoints
"A"(4)	4	16
"B"(3)	3	9
"C"(2)	3	6
"D"(1)	0	0
"F"(1)	1	1
"W"(1)	2	2
Sums	13	34

34/13 = a contextual research GPA of 2.62 and with a "2" in the target course (a "C").

With a contextual GPA for every student enrolled in the target course, we have an up-to-date measure of student ability (plus indirect measures of motivation, study habits, etc.) The only students without a contextual GPA would be those new freshmen with no prior college work and who enrolled only in the target course.

The correlation between the 4-pt. contextual GPAs and the research GPAs which include grades in the target courses is .98. The shape of the distribution of contextual GPAs should closely resemble the distribution for the research GPAs and it does (see Figure 6).

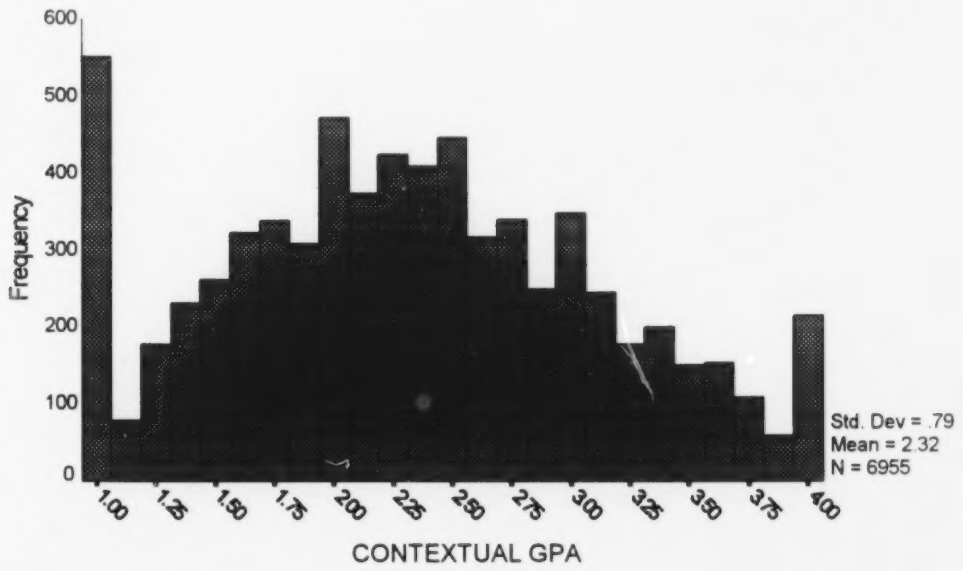

Figure 6. Histogram showing frequency of student contextual GPA.

We believe the contextual GPA to be the best covariate measure of student performance when evaluating differences in instructor grading patterns. It is definitely superior to an assessment test score because it includes more information about student performance and is apt to be more reliable because it is based upon extended behavior patterns.

By statistically equating students on contextual GPA through the use of analysis of covariance, or by examining its contribution as in multiple regression, any residual differences between instructor grading patterns are likely due to the instructor and not to initial student differences in academic ability levels.

We also believe the contextual GPA will be useful as a multiple measure in co/prerequisite research. By the time some students enroll in a target course, their assessment scores are dated, they may not have taken recommended preparatory skill courses, or they may be new students having no regular college GPA. By using the contextual GPA (or a conventional GPA computed in a contextual manner), the researcher should have a performance measure on nearly every student. Pending the outcome, a prerequisite for a course could be modified to read some score on the assessment instrument, completion of a preparatory skill course, or a particular college GPA on a specified number of units. Naturally, any published prerequisite GPA would refer to the regular GPA without contextual consideration. The contextual GPA is only used in the initial research.

Applying The Research GPA To Instructor Grading Variability Across The Curriculum

From the fall 1992 semester, we selected 19 general education courses with multiple sections that were offered during the day and routinely taught by at least two or more full-time, tenured instructors at ARC. The courses covered a span of four regular semesters. Different sections of the same course that were taught by the same instructor were combined to increase sample size for each faculty member. There were a total of 6,955 students and 71 instructors (69 unduplicated count).

We applied the 4-point contextual GPA for each student within a given course with a specific instructor as well as the 4-point research GPA for the instructor (i.e., the research GPA for the entire course per instructor). These results are presented in Table 2 (p. 15, 16). In examining the first course in Table 2 (Accounting 1A), you will note that instructor #1 had students who had a cumulative contextual GPA of 2.26. The research GPA, based only upon the grades in the target course, came in at 2.00. The difference between those two values (the gain or loss) is an average grade loss of -.26 with that instructor (course average minus contextual GPA). So, on average, the course was more difficult (i.e., more unsuccessful grade notations) than what these students had typically experienced. Notice that for Accounting 1A, all three instructors gave grades that resulted in grade loss for students. Now examine English 1A with its eight instructors. The gain/loss column (the differences) clearly show grading inconsistency. The biggest gain is with instructor #24 and the largest loss is with instructor #21. Finally, please examine the results for Statistics 1 with its three instructors. The absolute range between contextual GPA and instructor GPA is from -.54 to +.47, a difference of over one full research grade! Such large instructor grading variation within the same course is something that should not exist. Further, this magnitude of grading inconsistency renders validation of course placement by any type of student assessment as totally absurd. We have found that presenting data to faculty as we have organized it in Table 2, makes the issue abundantly clear.

Table 2

Student Contextual GPA's, Instructor GPA's, and Differences

by Course and by Instructor

Course	Instructor	Contextual GPA	Research GPA by Instructor	GPA Gain or Loss
Accounting 1A	1	2.26	2.00	-.26
	2	2.12	1.82	-.30
	3	2.59	2.05	-.54
Art 10	4	2.24	2.72	+.48
	5	2.48	2.88	+.40
Biology 25	6	2.74	2.54	-.20
	7	2.77	2.57	-.20
Business 18A	8	2.55	2.53	-.02
	9	2.34	1.90	-.44
Chemistry 1A	10	2.72	2.64	-.08
	11	2.75	2.55	-.20
	12	2.62	2.42	-.20
	13	2.69	2.28	-.41
Chemistry 2A	14	2.41	2.69	+.28
	15	2.62	2.27	-.35
	16	2.53	2.07	-.46
	17	2.65	2.39	-.26
English 1A	18	2.25	1.96	-.29
	19	2.34	2.75	+.41
	20	2.37	2.26	-.11
	21	2.32	1.71	-.61
	22	2.41	2.63	+.22
	23	2.52	2.81	+.29
	24	2.23	3.09	+.86
	25	2.53	2.37	-.16
English 58	26	1.95	1.77	-.18
	27	1.82	2.00	+.18
	28	2.09	1.57	-.52
	29	1.82	1.57	-.25
	30	2.06	2.16	+.10
	31	2.01	1.91	-.10
	32	2.03	1.55	-.48
	33	1.92	2.17	+.25
	34	2.14	2.13	-.01
English 271	35	2.14	2.40	+.26
	36	1.94	2.13	+.19
	37	2.09	2.09	+-.00

Table 2 (continued).

Course	Instructor	Contextual GPA	Research GPA by Instructor	GPA Gain or Loss
Health Ed 10	38	2.39	2.28	-.11
	39	2.28	2.43	+.15
	40	2.24	2.08	-.16
	41	2.40	2.52	+.12
History 17	42	2.20	1.57	-.63
	43	2.11	1.85	-.26
	44	2.22	1.87	-.35
	45	2.56	2.04	-.52
Math 51	46	2.23	1.77	-.46
	47	2.01	1.82	-.19
	48	1.96	1.97	+.01
Math 53	49	2.36	1.98	-.38
	50	2.31	1.69	-.62
	51	2.25	1.65	-.60
Music 6	52	2.43	3.06	+.63
	53	2.32	2.46	+.14
Philosophy 6	54	2.36	2.28	-.08
	55	2.37	1.96	-.41
Political Sci 1	56	2.16	1.85	-.31
	57	2.07	1.54	-.53
	58	2.42	1.94	-.48
	59	2.36	1.89	-.47
Psychology 1	60	2.07	1.95	-.12
	61	2.17	2.05	-.12
	62	2.19	2.39	+.20
	63	2.34	1.93	-.41
Speech 1	64	2.34	2.38	+.04
	65	2.47	2.48	+.01
	66	2.36	2.42	+.06
	67	2.46	2.91	+.45
	68	2.40	2.76	+.36
Statistics 1	69	2.38	1.88	-.50
	70	2.68	3.15	+.47
	71	2.69	2.15	-.54

Regression Analysis

To determine the actual contribution of "instructors" to the prediction of student grades in specific courses, we used step-wise multiple regression analysis. The criterion or dependent variable was research grade in the target course (where "A" = 4, "B" =3, "C" or CR =2, and "D", "F", "NC" or "W" =1). Cumulative 4-pt. contextual GPAs of students were entered first into the regression as a predictor variable and the R^2 value

noted. Next, the dummy codes for instructor were entered and the cumulative R^2 noted. The result was a unique change in R^2 which is the variance accounted for by knowledge of the instructor having adjusted for any differences in student contextual GPA. The complete results showing course, simple Pearson r, multiple R and R^2 gain by knowledge of instructor are shown in Table 3.

Table 3

Simple R's and Multiple R's Between Contextual GPA and Research Grade In Target Course Plus Gain by Knowledge of Instructor

Course	Total Students	Total Instructors	Simple R Contextual GPA With Course Research Grade	Multiple R Contextual GPA + Instr. With Course Research Grade	R^2 Gain with Instr.
Accounting 1A	208	3	.619	.628	.011
Art 10	286	2	.539	.540	.000
Biology 25	165	2	.678	.678	.000
Business 18A	220	2	.629	.669	.052***
Chemistry 1A	227	4	.625	.634	.011
Chemistry 2A	314	4	.561	.628	.079***
English 1A	585	8	.503	.625	.138***
English 58	433	9	.528	.585	.063***
English 271	140	3	.640	.647	.009
Health Ed 10	478	4	.680	.692	.016**
History 17	612	4	.583	.594	.013**
Math 51	219	3	.512	.533	.021*
Math 53	260	3	.572	.586	.016*
Music 6	351	2	.568	.600	.038***
Philosophy 6	149	2	.581	.595	.017
Political Sci 1	990	4	.586	.591	.006*
Psychology 1	757	4	.642	.671	.037***
Speech 1	291	5	.609	.627	.022*
Statistics 1	270	3	.506	.649	.165***
Totals	6,955	71			
		Median Value	.583	.627	.017
		Range	.503 to .680	.533 to .692	.000 to .165

* = $p < .05$
** = $p < .01$
*** = $p < .001$

The results in Table 3 parallel the findings in Table 2. We had made it difficult to establish wide grading variability because we used contextual GPAs (rather than assessment test scores) and a restricted 4-point research grade scale in the target course.

In spite of this, we did find some troublesome courses, but not wide grading inconsistency everywhere. Of course, this was a bit of good news!

In Table 3 in the R^2 Gain column, you will notice that the biggest trouble spots occur with the following courses: Business 18A, Chemistry 2A, English 1A, English 58, Music 6, Psychology 1, and the very worst problem, Statistics 1. All of these courses had instructors which accounted for 3% or more of the grading variance not attributable to differences in student grade histories. With Statistics 1, knowledge of instructor accounted for 16.5% of the variance in grading. Overall, the median R^2 gain by knowledge of instructor was 1.7%.

In Table 3, note that the simple correlations between contextual GPAs and course grades have a median value of .583 which, from our practice, is globally higher than correlation coefficients between assessment test scores and grades. Frankly stated, college GPAs are generally better predictors of target course grades than are assessment test scores. In our judgment, cumulative college GPA should be given more official status as a multiple measure when judging student qualifications for course placement.

But Before You Trash The Test

It has been established that wide differences in instructor grading practices can create havoc with validating any assessment/placement procedure. The sensible thing to do is to start a movement toward rectifying grading inconsistency. The political hammer is "no consistency-- no placement."

Should you wish to examine what a validity coefficient between assessment test scores and grades might be under the most favorable of circumstances (i.e., little or no instructor grading differences), we offer the following recommendations.

1. Compute a part (or partial) correlation between assessment test scores and grades having removed the effects of instructor GPA from grades (part) or from grades and assessment scores (partial). The part procedure is outlined as Option D, page 21.10 in the yellow covered Matriculation Evaluation: Phase III Local Research Options, June 1992. Note that a part or partial correlation value may sometimes be lower than the original correlation between assessment scores and grades. This is a function of the interrelationships existing between the variables. For example, if the correlation between GPAs for instructors and student assessment scores is high positive, the part or partial correlation between assessment scores and grades will be lower than the original correlation.

2. A more straightforward approach is to convert grades for all students enrolled with the same instructor into standardized Z scores (the student's coded grade value minus the mean GPA for a specific instructor all divided by the standard deviation of grades for that instructor). Converting student grades into respective Z scores within each instructor makes the mean of all Z scores (for

each instructor) equal to 0.0. The student Z scores are now relative performances. A Z of 1.00 is interpreted in the same way across all instructors, namely that such a student scored one standard deviation above the mean grade (GPA) for the course with that instructor. It may be that a grade of "B" with one instructor could convert to a Z of 1.00, but the same grade if earned from another instructor could convert to a Z of 2.0.

Once the Z conversions are completed for each instructor within the same course, a Pearson r correlation can be computed between student assessment test scores and respective Z scores that are based upon grades. The result will give the correlation between assessment scores and relative course performances with the effects of instructor grading inconsistency removed.

Please remember that such procedures as outlined above will not validate the traditional use of an assessment instrument. Rather, it will help to gain some information on what validity could be under ideal circumstances. This may help to remove doubts about an assessment test and place it upon inconsistency of grading if that is where it truly belongs.